Well-ordered Collaboration Structures of Co-Author Pairs in Journals

HILDRUN KRETSCHMER AND THEO KRETSCHMER

In single-authored bibliographies only single scientist distribution can be found. But in multi-authored bibliographies single scientists distribution, pairs distribution, triples distribution, etc., can be presented. Whereas regarding Lotka´s law single scientists P distribution (both in single-authored and in multi-authored bibliographies) is of interest, in the future pairs P, Q distribution, triples P, Q, R distribution, etc. should be considered. Starting with pair distribution, the following question arises in the present paper: Is there also any regularity or well-ordered structure for the distribution of co-author pairs in journals in analogy to Lotka's law for the distribution of single authors? Usually, in information science "laws" or "regularities" (for example Lotka's law) are mathematical descriptions of observed data in form of functions; however explanations of these phenomena are mostly missing. By contrast, in this paper the derivation of a formula for describing the distribution of the number of co-author pairs will be presented based on well-known regularities in socio-psychology or sociology in conjunction with the Gestalt theory as explanation for well-ordered collaboration structures and production of scientific literature, as well as derivations from Lotka's law. The assumed regularities for the distribution of co-author pairs in journals could be shown in the co-authorship data (1980-1998) of the journals Science, Nature, Proc Nat Acad Sci USA and Phys Rev B Condensed Matter.

Introduction

Over the time collaboration is increasing in science and in technology. Usually, the bibliometric method for the study of collaboration is the investigation of co-authorships. Collaboration between countries, collaboration between institutions, or collaboration between individual scientists is covered in the literature [1; 2; 3; 4; 12; 13; 14; 15; 17; 19; 22]. Single authored bibliographies are changing to multi-authored bibliographies with increasing number of co-authors per paper.

In single authored bibliographies only single scientist distribution can be found. But in multi-authored bibliographies single scientists distribution, pairs distribution, triples distribution, etc., can be presented. Several methods are possible for counting single scientists, pairs, triples etc. These are:

- Normal count procedure, in which each co-author (single scientists P counting like in Lotka's law) or each pair (pairs P, Q counting), etc., receives full credit,
- Weighted (or fractional) counting, in which each co-author of m co-authors per paper receives $1/m^{th}$ credit.

The famous Lotka's law discovered in 1926 keeps on fascinating scientists from all fields.

Lotka´s law based on single scientists P counting is valid for:

- single authored bibliographies, and,
- under the condition of normal count procedure, also for multi-authored bibliographies

But:

- Rousseau [16] showed that weighted (or fractional) counting leads to a breakdown of Lotka´s regularity
- In addition, it is shown empirically that, even when one uses the normal or total counting procedure, Lotka's law breaks down when articles with a large, i.e. more than hundred, number of authors are included in the bibliography [9].

Whereas regarding Lotka's law single scientists P distribution (both in single-authored and in multi-authored bibliographies) is of interest, in the future pair P, Q distribution, triple P, Q, R distribution, and so on, should be considered using normal counting procedure.

Starting with pair distribution, the following question arises in the present paper: Is there also any regularity for the distribution of co-author pairs in journals in analogy to Lotka's law for the distribution of single authors?

Usually, in information science "laws" or "regularities" (for example Lotka's law) there are mathematical descriptions of observed data in form of functions; however explanations of these phenomena are mostly missing. By contrast, in this paper the derivation of a formula for describing the distribution of the number of co-author pairs will be presented based on well-known regularities in socio-psychology or sociology in conjunction with the Gestalt theory as explanation for well-ordered collaboration structures and production of scientific literature, as well as derivations from Lotka's law. A modified presentation has already been published in [7] and in [10]. The model was tested with co-authorship data of the journal Science. There came up the question, whether these findings are also valid for other journals. Therefore, the new results of the study of the co-authorship data of three additional journals will be shown in this paper:

- Nature
- Proc Nat Acad Sci USA
- Phys Rev B Condensed Matter.

The formula for describing the distribution of the number of co-author pairs P, Q between authors with i publications per author and authors with j publications per author (N_{ij}) will be derived by a combination of:

- The possible distribution of the number of pairs N_{ij} under the condition of derivation alone from the distributions of the marginal sums N_i and N_j. For explanation: [18] showed the number of collaborators N_i (or N_j respectively) is distributed in the same way as the total number of publications of all authors with i publications per author, i. e. there is a connection with Lotka's law regarding this possible distribution.
- The measure of the influence of social structures on the distribution of the number of co-authorship pairs N_{ij} regardless of the marginal sums N_i and N_j. Before derivation of the formula the methods are presented for counting N_i and N_{ij}.

Methods for Counting N_i and N_{ij}

Different versions of counting pairs P, Q are possible. Two of them will be presented below:
Given is a bibliography (partly represented, names of authors A, B...)

1.	A	4.	D, A, F	7.	H, G
2.	B	5.	C	8.	H, G, A
3.	D, E	6.	G, H	etc.	

The number of publications i per author P is determined by resorting to the 'normal count procedure'. Each time the name of an author appears, it is counted (e. g. three times: once in the first article, and once each in the 4th and 8th article).

It should be noted here that the term 'article' is used with reference to a work or a paper written by one or jointly by several authors, (see some articles in the bibliography). By contrast, the term 'publication' refers to persons.

First version:
Pairs P, Q are counted analogue to single authors P in Lotka's law. The number of publications n_c per pair P, Q is determined (D, E one time; D, A one time; D, F one time; A, F one time; G, H one time; H, A one time; G, A one time; H, G two times).

Second version:
Pairs P, Q are counted under the condition of both the first authors P count (i) and the second authors Q count (j), i. e. the authors are respectively ordered according to i or j (cf. Table 1).

In the present paper the symmetrical form of matrix 1 will be studied, see Table 2. In the symmetrical matrix, one can determine for each author P the number of his collaborators N_P as well as the total sum of his relationships through co-authorships C_P.

Table 1: Matrix 1

i / j		1				2	3		
		B	C	E	F	D	A	G	H
1	B								
	C								
	E								
	F								
2	D			1	1		1		
3	A				1				
	G						1		1
	H						1	2	

The number of collaborators N_P is the number of pairs assigned to the author P in the corresponding row of the matrix. The number of relationships through co-authorships can differ between the pairs. For example, the author P=H has two collaborators ($N_H = 2$), namely Q_1=A and Q_2=G. While the author H has one relationship with the collaborator A through co-authorship, there are three relationships with the collaborator G. The sum of the relationships through co-authorships for the author P=H is, thus, four (C_H=4).

Table 2: Symmetrical form of matrix 1

i / j		1				2	3				
		B	C	E	F	D	A	G	H	Np	Cp
1	B										
	C										
	E					1				1	1
	F					1	1			2	2
2	D			1	1		1			3	3
3	A				1	1		1	1	4	4
	G						1		3	2	4
	H						1	3		2	4
									SUM	14	18

Two different matrices can be derived from matrix 2:

- Table 3 is the representation of the number of pairs N_{ij} with authors who have i publications per author, with authors who have j publications per author, included in the bibliography. A_i is the number of authors with i publications per author. $N_i = \Sigma_j N_{ij}$ is the number of collaborators of all authors with i publications per author. $R_i = k = N_i/A_i$ is the average number of collaborators per author. A former investigation on the topic collaboration and Lotka's law proposed the average number of collaborators per author k to predict the productivity strata [18]. An experiment was introduced in Lotka's law with the new variable (k).
- Table 4 is the representation of the number of co-authorship relations C_{ij} between authors with i publications per author, and authors with j publications per author. The matrix C_{ij} was the pre-requisite for earlier studies ([5] and 2002 for more details). Because of the possible fluctuation at that time, however, a classification of the data i and j was done corresponding to the logarithm resulting in a matrix C_{XY}. The structure of this matrix C_{XY}, also as of the corresponding relative values, has been described. The results of this study have been taken as a pre-condition for the present work.

Table 3: Matrix of N_{ij}

i / j	1	2	3	N_i	A_i	R_i
1	0	2	1	3	4	0.75
2	2	0	1	3	1	3
3	1	1	6	8	3	2.67
			SUM	14	8	

Table 4: Matrix of C_{ij}

i / j	1	2	3	C_i
1	0	2	1	3
2	2	0	1	3
3	1	1	10	12
			SUM	18

Hypotheses

The hypotheses were derived from the following considerations:

- How would the distribution of the number of pairs looks like, if it is derived alone from the marginal sum N_i, that is from the number of collaborators of all authors with i publications per author?
- Measurement of the influence of characteristic features of social structures on the distribution of N_{ij} regardless of the marginal sum $N_{i,}$
- Derivation of a formula for describing the distribution of the number of pairs N_{ij} including both the distribution of the marginal sums and the characteristic features of social structures.

Hypothesis 1

There is a possible distribution of the number of pairs N_{ij} derived alone from the marginal sums N_i and $N_{j,}$ under the condition there is not any social influence on collaboration. In her attempt to give an explanation [18] showed in her example, that the number of collaborators N_i is distributed in the same way as the total number of publications of all authors with i publications per author (T_i):

$$T_i = i\,A_i \qquad (1)$$

This means, that the marginal sums N_i (or N_j respectively) should be distributed according to an inverse power function in line with Lotka's law, however, with a different parameter:

Ni = constant/ia (2)

Because of the symmetry of the matrix, both the distributions of the marginal sums of the matrix are equal (row = column).

Assuming the condition, that the productivity of the authors has no social influence whatsoever, which author collaborates with which other author, the distribution of the number of pairs could be determined within the matrix solely on the basis of the marginal sums:

$N'_{ij} = N_i \cdot N_j/\Sigma_i N_i$ (3)

Under the condition the formula (2) is valid there could be a relationship here with Lotka's law:

$N''_{ij} = \text{constant}/(i \cdot j)^a$ (4)

Hypothesis 2

There is a possibility to measure the influence of social structures on the distribution of the number of co-authorship pairs N_{ij} regardless of the marginal sums N_i and N_j as follows:

Explanation: To determine whether the distribution of data within a matrix shows additional characteristic features, which have raised independent of the distribution of the marginal sums, the homophylic index H_{ij} can be used.

In some sociological studies of interpersonal relations in social networks of men [21], this special homophylic index is used. That index provides information on the factor, by which the observed frequency in a cell of a matrix deviates from the occupancy of this cell that would otherwise be expected in case of statistical independence from characteristics. In order to calculate this index, we have to covert the matrix of observed frequencies N_{ij} into a new matrix using geometric mean. The special homophylic index H_{ij} is defined as:

$$H_{ij}=N_{ij} \cdot \frac{G}{G_i \cdot G_j} \qquad (5)$$

where G - geometric mean of all matrix data

G_i - geometric mean of the data in row i

G_j - geometric mean of the data in column j.

Under the condition the distribution of data within the matrix is determined by the distribution of the marginal sums N_i and N_j. The resulting homophylic index H'_{ij} is equal to 1 in each cell of the matrix. Thus, H'_{ij} is valid starting from the distribution of N'_{ij} produced according to formula (3).

On the other hand, if the distribution of the actually observed N_{ij} is determined beyond the marginal sums through the structures, which are in general valid in social structures. Such structures must be provable in the distribution of $H_{ij,}$ and the distribution of H_{ij} must be different from that of H'_{ij}.

How do these social structures appear? Following the general characteristic features of social structures are represented, while discussing the structural characteristics of interpersonal relations in social networks. Reference shall be made to one of [21]'s work rather than to the many studies conducted and contained in the literature. As a result, a definite structure can be identified that underlies a great number of social processes of a distributive character, such as the spreading of diseases, the propagation of information, the change of views or the distribution of innovations. A generalization of this structure reveals three pivotal aspects:

1. Over-coincidental similarity among persons in contact with each other ("Birds of a feather flock together");
2. Decrease of interpersonal relations with declining similarity;
3. Emergence of the 'edge effect' (cf. below).

The author illustrates these three aspects on the basis of an empirical example investigated by using the homophylic index ([21], p. 35). Independently of whether socio-demographic features, socio-structural characteristics or general approaches are taken into account or not, it has repeatedly been shown that persons with social contacts reveal greater characteristic similarities than it could have expected from persons with accidental associations. Relations may be qualified as friendships, marriages, professional contacts, collaboration or other types of relationship.

Wolf [21], in one of his empirical examples, studied similarity underlying relations of friendship due to common education. It unequivocally appears, that preferably become friends those persons, who had achieved the same level of education. These data of the same level of education can also be used to observe the edge effect. This term designates the more pronounced similarity of friendly couples that is observable at the edges of status features (referring to persons at the lowest, as well as to those at the highest levels of education). Using the data file of [21] it is possible to identify four-times-higher relations between high-school leavers and university graduates than it would be expected at a fortuitous choice of friends. The tendency to choose status-homogeneous friends is less clearly perceptible with persons having medium-level school degrees. Resultant at the same level of education arose a U-curve of data.

In the above mentioned papers [5] and [2] it was shown in more details, that these general basic structures of social networks are also valid in co-authorship networks. This was shown by analyzing the relationships through co-authorships (Matrix of C_{ij}). The present paper will further examine, whether these basic structures also appear in the analysis of pairs (Matrix of N_{ij}) because N_{ij} and C_{ij} can vary independently of one another.

Hypothesis 3

The Derivation of a formula for describing the distribution of the number of pairs N_{ij} is possible as follows:

The first pivotal aspect of the interpersonal relations in social networks, i. e. the over-coincidental similarity among persons in contact with each other, the well known proverb "Birds of a feather flock together" could be conveniently integrated into the theory, together with the empirical results published (Kretschmer 2002). In the literature of sociology far less evidence is found, however, for the opposite saying "Opposites attract", although several efforts have also been put into proving its correctness, just think of [20].

By contrast, the author (Kretschmer 2002) has made an attempt to suggest that both opposing proverbs should only be perceived as the conspicuously visible state of a holistic process occasioned by conditions to which the system under study was subjected during the investigation. Moreover, the same applies to both opposing views of U-curves, i. e. with edge effect on the one the side, and the reverse case, on the other hand.

In this context it could be shown (Kretschmer 2002), that four factors that can vary independently of one another, influence the authors, who makes a co-authorship with whom:

- Dissimilarity ($A=|\log i - \log j|$) regarding their productivity;
- Complementary to that similarity ($A_{COMPLEMENT} = A_{max}+A_{min} - A$)
- (Both the first two factors concern “Birds of a feather flock together” or “Opposites attract”);
- Another independent, influencing factor is the sum $B = \log i + \log j$;
- Complementary to this ($B_{COMPLEMENT} = B_{max}+B_{min} - B$).

For example, in the case of A= 0= constant on one hand the authors could be having low productivity with B_{min}, or on the other hand highly productive with B_{max} or also authors with any average productivity. Here: $A_{min} = |\log 1 - \log 1| = 0$ and $B_{min} = \log 1 + \log 1 = 0$.

Both the last two factors concern the edge effect, visible through the U-curve or the reverse case.

(Note: In the former publications a classification of the data i and j was done corresponding to the logarithm. Class X contained those authors with 1 publication per author, class X=2 authors with 2-3 publications, X=3 authors with 4-7 publications, X=4 authors with 8-15 publications, X=5 authors with 16 and more publications; by analogy the same applied to Y; following $A=|X-Y|$ and $B=X+Y$.)

The four influencing factors, that are independent of one another, have been included in the formula for the description of co-authorship structures. Similar to the formula of earlier papers [8] and Kretschmer 2002, it is assumed here for describing the distribution of H_{ij} approximately:

$$H_{Tij} = \text{const} \cdot (A + 1)^{\tau} (A_{COMPLEMENT}+1)^{\omega} \cdot (B+1)^{\xi} \cdot (B_{COMPLEMENT}+1)^{\psi} \quad (6)$$

with $A = |\log i - \log j|$ and $B = \log i + \log j$ and with the parameters τ, ω, ξ, and ψ.

It has been mentioned an earlier paper, that the values of the individual parameters depend on the interplay of the influencing factors and their complements. For example, the lengthening of the environmental conditions can have an effect on this interplay, and hence can effect a change of the parameter. In this context, different types of collaboration patterns have been presented. How far the influence of the similarity of the authors regarding their productivity retreats for the benefit of the effect of dissimilarity and vice versa, depends on the ratio of the parameters τ and ω. These changes take place continuously. The same is also true of the parameters ξ and ψ with respect to the edge effect.

The logarithm of the maximum possible number of publications of an author corresponds to A_{max}. Thereby, $B_{max}= 2 \cdot A_{max}$. It is possible to lay down a specific value as standard for such studies, which does not vary depending upon the given sample. It is assumed that the maximum possible number of publications of an author is equal to 1000, i. e.

$$A_{max}=\log 1000=3$$

It is assumed, that the formula given above for the distribution of H_{ij} is also suitable for describing the number of pairs N_{ij}, because N_{ij} is influenced on one hand by the social structures, and, independent of that, on the other hand, also by the distribution of the marginal sums N_i and N_j.

Since the distribution of N''_{ij} can also approximately be described by the product of the four factors of the formula H_{Tij} (Under the condition of logarithmic presentation after regression analysis the following correlation could be found R=0.9999, F=526,038 with n=496). It is assumed that while changing the values of the parameters, the distribution of the number of co-author pairs additionally influenced by the social structures can be described by the following formula:

$$N_{Tij} = const \cdot (A + 1)^{\alpha} \cdot (A_{COMPLEMENT}+1)^{\beta} \cdot (B+1)^{\gamma} \cdot (B_{COMPLEMENT}+1)^{\delta} \quad (7)$$

with $A = |\log i - \log j|$ and $B = \log i + \log j$ and with the parameters α, β, γ, and δ

Data and Results: An Overview

The next paragraph shows an empirical example how to understand the theoretical assumptions. In this line the empirical results of the studied co-authorship data obtained from the journal Science are visualized:

Firstly the distribution of the number of single authors P with i publications per author (Reflection of Lotka's law, two-dimensional) is compared with the distribution of the number of pairs P, Q with i and j publications per author N_{ij} (three-dimensional). Fig. 1 with detailed explanation is attached in the next paragraph.

Secondly the theoretical assumed influence of the social structure is visualized by the empirical data obtained by the homophylic index H_{ij}. Figure 2 clearly shows, that the general social structures – independent of the marginal sums N_i and N_j – have exercised an influence on the distribution of N_{ij}.

Thirdly the empirical distribution of pairs N_{ij} is compared in Fig. 3 with two different distributions of theoretical values:

- First theoretical distribution: the theoretical values N''_{ij} are determined through regression analysis under consideration of N_i and N_j only, formula (4) above.
- Second theoretical distribution: the theoretical values N_{Tij} are determined through regression analysis under consideration of the derived formula (7) above.

The empirical distribution is rather equal to the second theoretical distribution but different from the first one. This is the proof for the influence of social structures on the distribution of N_{ij}.

After this paragraph another will follow with presentations of the empirical distributions of N_{ij} obtained from three other journals in comparison with theoretical distributions mentioned above as "second theoretical distribution".

Data and Results of the Journal Science

The articles from 1980-1998 of the journal Science were studied with 47,117 authors and the total sum of N_{ij}=418,458 co-author pairs.

Figure 1 shows the distribution of the number of authors A_i with i publications per author contrasted with the distribution of the number of co-authorship couples (pairs) N_{ij} for clarity's sake and optimum visualization restricted to authors with at most ten articles i(i=1,2…10).

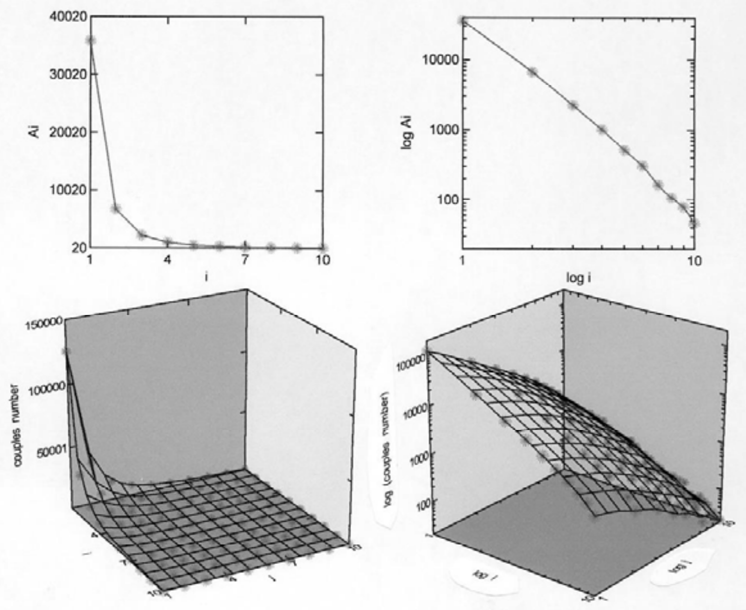

Fig. 1: Distribution of the number of authors A_i with i publications per author contrasted with the distribution of the number of co-authorship couples N_{ij} with i(i=1,2…10).
The upper row reflects the Lotka's law. In this row, the distribution of A_i is given on the left, and on the right the corresponding double logarithmic presentation.
In the lower row, the distribution of the number of co-authorship couples N_{ij} is given on the left, and the corresponding triple logarithmic presentation on the right.

The upper row reflects the Lotka's law. In this row, the distribution of A_i is given on the left, and on the right the corresponding double logarithmic presentation (R=0.997, F=1325.6, P≈0 with n=10). In the lower row, the distribution of the number of co-authorship pairs N_{ij} (pairs' number) is given on the left, and the corresponding triple logarithmic presentation on the right. The number of triples N_{ij} or log N_{ij} is respectively plotted at the Z-axis. Under the condition of logarithmic presentation after regression analysis the following correlation could be found:

- R=0.998, F=3317.5, P≈0 with n=55 values (because of symmetry, n corresponds to the half matrix only, plus main diagonal) for log N_{ij}.
- R=0.993, F=615.6, P≈0 with n=10 are valid regarding the distribution of log N_i according to an inverse power function, see formula (2).

Under the condition the data were cut off after i=15 about 99.9% of the total number of authors are included:

- R=0.996, F=1586.9, P≈0, n=15 regarding the distribution of log A_i
- R=0.993, F=899.5, P≈0, n=15 regarding the distribution of log N_i
- R=0.995, F=2881.7, P≈0, n=120 regarding the distribution of log N_{ij}

Under the condition of using all data the following correlations could be found:

- R=0.978, F=619.7, P≈0 with n=30 regarding the distribution of log A_i in comparison with Lotka's law;
- R=0.934, F=194.5 , P≈0 with n=30 regarding the distribution of log N_i;
- R=0.980, F=1668.0, P≈0 with n=280 regarding the distribution of log N_{ij} (Here the five authors with the highest number of publications are excluded because of matrix size).

Figure 2 shows the homophylic index, i (i=1, 2…0). The distribution of H'_{ij} resembles a plane surface because the homophylic index H'_{ij} is equal to 1 in each cell of the matrix, see the four cases. The figures on the right are rotated by 90°. In the upper row, the distribution of H'_{ij} is compared with the distribution of the theoretical values $H_{Tij.}$ These theoretical values are determined through regression analysis of the empirical values H_{ij} regarding the formula (6) in the hypothesis (R=0.840, F=30.2, P≈0 with n=55 for log H_{ij}). In the lower row, the distribution of H'_{ij} is compared with the distribution of the empirical values H_{ij}.

Fig. 2 clearly shows that the general social structures – independent of the marginal sums N_i – have exercised an influence on the distribution of N_{ij}.

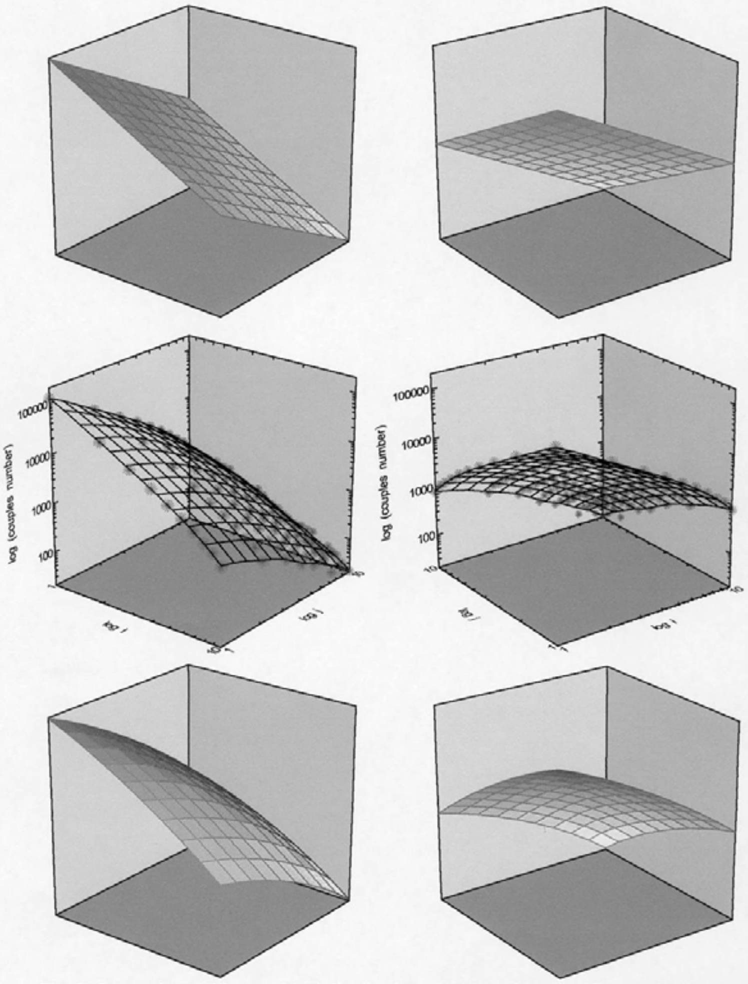

Fig. 3: Comparison of the empirical distribution of log N_{ij} with two different distributions of theoretical values. The figures on the right are rotated by 90°
In the first row, the distribution of theoretical values is shown under the condition:
log N''_{ij} =constant+a•(log i +log j). The 2nd row shows the empirical distribution of log N_{ij}.
In the 3rd row the theoretical values are shown under the condition:
$\log N_{Tij} = \text{const} + \alpha \bullet \log(A + 1) + \beta \bullet (A_{max}-A+1) + \gamma \bullet \log(B+1) + \delta \bullet \log(B_{max}-B+1)$

Fig. 3 shows the comparison of the empirical distribution of log N_{ij} with two different distributions of theoretical values. The figures on the right are rotated by 90°. In the first row, the distribution of theoretical values is shown under the condition

$$\log N''_{ij} = \text{constant} + a' \cdot (\log i + \log j) \quad (8)$$

The second row shows the empirical distribution of log N_{ij}, whereas in the third row the theoretical values are shown under the condition:

$$\log N_{Tij} = \text{const} + \alpha \cdot \log(A + 1) + \beta \cdot \log(A_{COMPLEMENT}+1) + \gamma \cdot \log(B+1) + \delta \cdot \log(B_{COMPLEMENT}+1) \quad (9)$$

R=0.988 and F=2178.3, P≈0 with n=55 after regression analysis of the empirical distribution of log N_{ij} with the theoretical pattern in the first row.
R=0.998 and F=3317.5, P≈0 with n=55 after regression analysis of the empirical distribution of log N_{ij} with the theoretical pattern in the third row.

Data and Results of the journals Nature, Proc Nat Acad Sci USA and Phys Rev B Condensed Matter

Articles of the following journals are studied from 1980-1998:

- *Nature* with 52,838 authors and the total sum of N_{ij}=581,698 co-author pairs;
- P*roc Nat Acad Sci USA* with 79,877 authors and the total sum of N_{ij}=704,032 co-author pairs;
- *Phys Rev B Condensed Matter* with 46,232 authors and the total sum of N_{ij}=544,006 co-author pairs.

The figures 4-6 show the empirical distributions of log N_{ij} plotted on the theoretical distributions obtained after regression analyses. The figures on the right are rotated by 90°. In analogy to the study of the co-authorship data in the journal Science above in the first rows the distributions are shown under the condition the data were cut off after i=10.

Following in the second row under the condition of cut off after i=15 (in the average about 99% of the authors are included) and in the third row under the condition of cut off after i= 31 (because of matrix size only a very few of the highest productive authors are not included).

Under the condition of logarithmic presentation after regression analysis the following correlations between theoretical and empirical values of log N_{ij} could be found for the three journals:

Nature (Fig. 4)

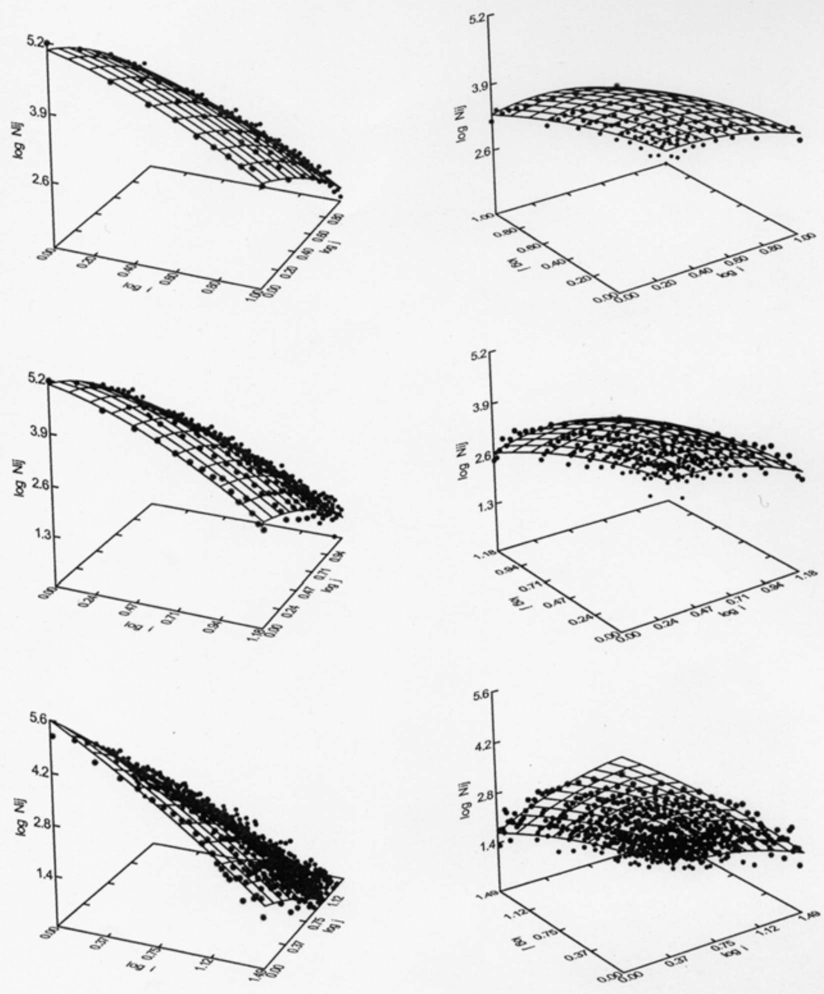

Fig. 4: Comparison of the empirical distributions of log N_{ij} in the journal *Nature* with the theoretical distributions after regression analysis.

Under the condition the data were cut off after i=10 the following correlation could be found:

- R=0.997, F=2105.30, P≈0 with n=53 values (because of symmetry, n corresponds to the half matrix only, plus main diagonal) for log N_{ij}.

Under the condition the data were cut off after i=15:

- R=0.992, F=1757.65, P≈0, n=117 regarding the distribution of log Nij.

Under the condition the data were cut off after i=31:

- R=0.972, F=1427.48, P≈0 with n=337 regarding the distribution of log N_{ij} (Here very few authors with the highest number of publications are excluded because of matrix size).

Proc Nat Acad Sci USA (Fig. 5)
Under the condition the data were cut off after i=10:

- R=0.999, F=7346.76, P≈0 with n=53 values for log N_{ij}.

Under the condition the data were cut off after i=15:

- R=0.997, F=4767.5, P≈0, n=118 regarding the distribution of log $N_{ij.}$

Under the condition the data were cut off after i=31 the following correlations could be found:

- R=0.976, F=2256.97, P≈0 with n=459 regarding the distribution of log $N_{ij.}$

Phys Rev B Condensed Matter (Fig. 6)
Under the condition the data were cut off after i=10:

- R=0.999, F=4202.29, P≈0 with n=53 values for log N_{ij}.

Under the condition the data were cut off after i=15:

- R=0.999, F=6559.79, P≈0, n=118 regarding the distribution of log N_{ij}

Under the condition the data were cut off after i=31 the following correlations could be found:

- R=0.988, F=5161.86, P≈0 with n=496 regarding the distribution of log $N_{ij.}$

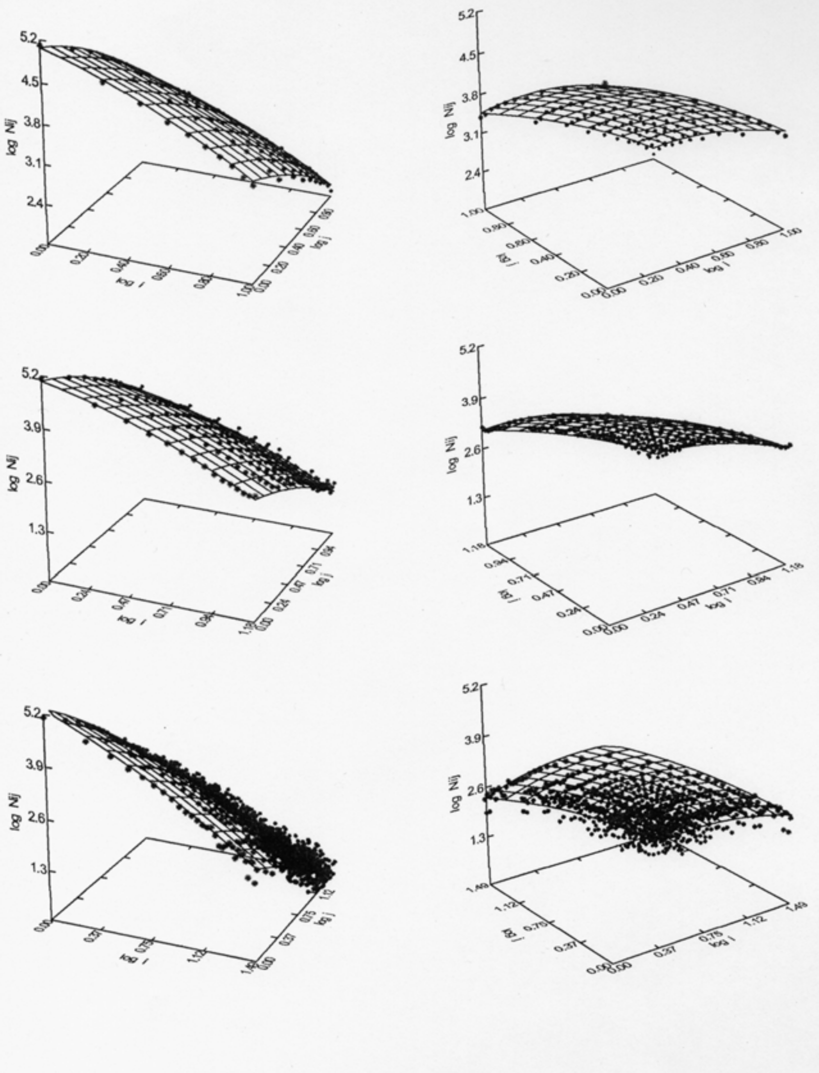

Fig. 5: Comparison of the empirical distributions of log N_{ij} in the journal *Proc Nat Acad Sci USA* with the theoretical distributions after regression analysis.

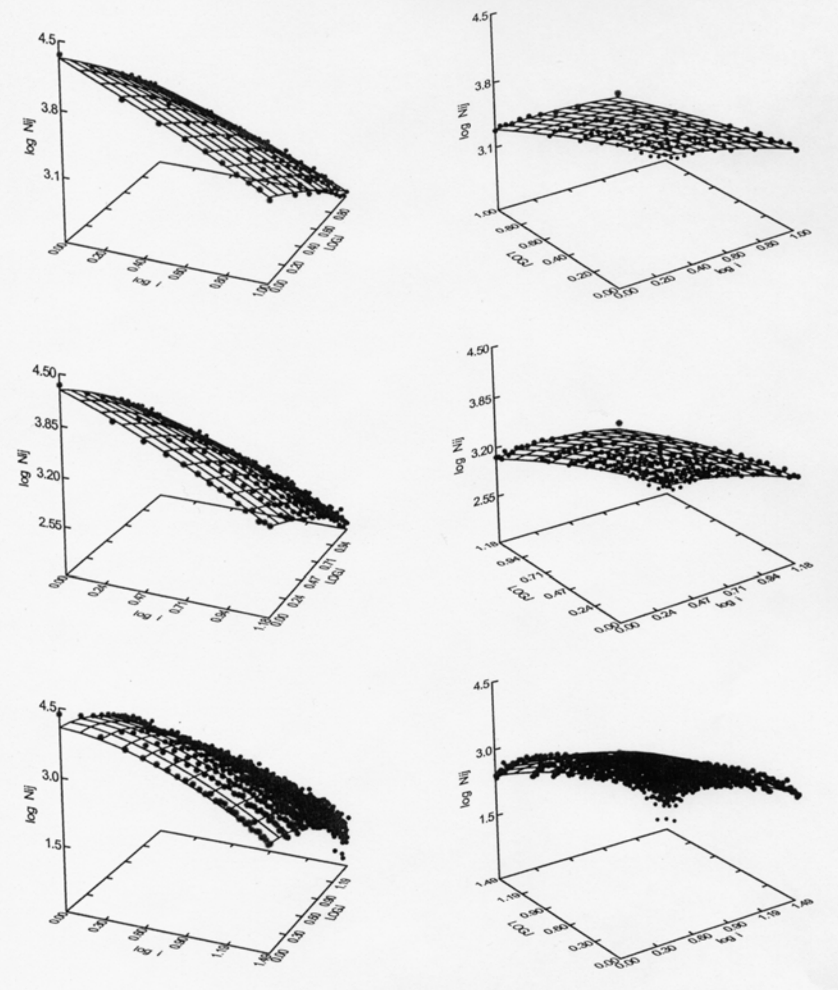

Fig. 6: Comparison of the empirical distributions of log N_{ij} in the journal *Phys Rev B Condensed Matter* with the theoretical distributions after regression analysis.

Concluding Remarks

In correspondence with [18] the number of collaborators N_i is distributed according to an inverse power function in line with Lotka's law. In this line it is also assumed, that the number of pairs N_{ij} is influenced by the distribution of the marginal sums N_i and N_j, i. e. by a deriviation of Lotka's Law:

$$N''_{ij} = \text{constant}/(i \cdot j)^a \quad (4)$$

There is the assumption the distribution of co-author pairs in journals on the third dimension is independently influenced by:

- Lotka's distribution mentioned above and;
- The general characteristic features in social systems.

The social influence is measured by:

$$H_{Tij} = \text{const} \cdot (A + 1)^{\tau} \cdot (A_{COMPLEMENT}+1)^{\omega} \cdot (B+1)^{\xi} \cdot (B_{COMPLEMENT}+1)^{\psi} \quad (6)$$

with $A = |\log i - \log j|$ and $B = \log i + \log j$ and with the parameters τ, ω, ξ, and ψ.

This formula was established on the basis of general knowledge about rules in social networks related to "Who is in contact with whom" and was published in [5], 1999 and in Kretschmer, 2002 in a more detailed fashion. The distribution of N''_{ij} can approximately be described by the product of the four factors of the formula H_{Tij}.

Thus, there is the assumption that while changing the values of the parameters in the formula (6), the distribution of the number N_{ij} of co-author pairs P,Q with i publications per first author P and with j publications per second author Q influenced both by Lotka's Law and by the social structures can be described by the following formula:

$$N_{Tij} = \text{const} \cdot (A + 1)^{\alpha} \cdot (A_{COMPLEMENT}+1)^{\beta} \cdot (B+1)^{\gamma} \cdot (B_{COMPLEMENT}+1)^{\delta} \quad (7)$$

with $A = |\log i - \log j|$ and $B = \log i + \log j$ and with the parameters α, β, γ, and δ

The presented results are findings in the journals Science, Nature, Proc Nat Acad Sci USA and Phys Rev B Condensed Matter with a large amount of data. There is the question if these rules are valid also in other journals or because of the necessary data extent only in combinations of several journals.

Acknowledgments

The author wishes to thank Holger Heitsch and Jan Johannes for programming.

References

1 BEAVER, D. DE B. (2001). Reflections on scientific collaborations (and its study): Past, present and prospective. *Scientometrics*, 52, 365-377.

2 GLÄNZEL, W. (2002). Co-authorship patterns and trends in the sciences (1980-1998). A bibliometric study with implications for database indexing and search strategies. *Library Trends*, 50, 461-473.

3 GLÄNZEL, W. & SCHUBERT A. (2004). Analyzing scientific networks through co-authorship. In H. F. Moed [et al.] (ed.), *Handbook of Quantitative Science and Technology Research* (pp 257-276). Dordrecht [u.a.]: Kluwer Academic Publishers.

4 GLÄNZEL, W. & DE LANGE C. (1997). Modeling and Measuring Multilateral Co-authorship in International Scientific Collaboration. Part II. A comparative study on the extent and change of international scientific collaboration links. *Scientometrics,* 40, 605-626.

5 KRETSCHMER, H. (1999). A New Model of Scientific Collaboration. Part I: Types of Two-Dimensional and Three-Dimensional Collaboration Patterns. *Scientometrics,* 46,.3, 501-518.

6 KRETSCHMER, H. (2000). Similarities and Dissimilarities in Co-authorship Networks; Gestalt Theory as Explanation for Well-ordered Collaboration Structures and Production of Scientific Literature. *Library Trends,* 50, 3, 474-497.

7 KRETSCHMER, H. (2001). Distribution of Co-Author Couples in Journals: "Continuation" of Lotka's Law on the 3rd Dimension. In *Proceedings of the 8th International Conference on Scientometrics & Informetrics, Sydney, Australia, July 16-20.* Edited by M. Davis & S. Wilson. Bibliometric & Informetric Research Group (BIRG (pp. 317-326). Sydney, Australia: UNSW.

8 KRETSCHMER, H., LIANG, L. & TUNDRA, R. (2001). Chinese-Indian-German collaboration results that provided the impetus for the foundation of COLLNET. *Scientometrics*, 52, 3, 445-456.

9 KRETSCHMER, H. & ROUSSEAU, R. (2001). Author Inflation Leads to a Breakdown of Lotka's Law. *Journal of the American Society for Information Science and Technology*, 52, 610-614.

10 KRETSCHMER, H. & KRETSCHMER, T. (2004). Comparison of rules in bibliographic and in web networks. In T.A.V. Murthy, S.M. Salgar, G. Makhdumi, P. Pichappan, Y. Patel and J. K. Vijayakumar (Ed.). *Proceedings of the Second International Caliber 2004: Road Map to New Generation of Libraries Using Emerging Technologies. February 11-13, 2004, New Delhi, India* (pp. 470-486). Ahmedabad: INFLIBNET Centre.

11 LOTKA, A. J. (1926). The frequency distribution of scientific production. *Journal of the Washington Academy of Science,* 16, 317-323.

12 LUUKKONEN, T., PERSSON, O. & SILVERTSE, G. (1992). Understanding patterns of international scientific collaboration. *Science, Technology and Human Values*, 17, 101-126

13 MIQUEL, J. F. & OKUBO, Y. (1994). Structure of International Collaboration in Science, Part II : Comparisons of Profiles in Countries using a Link Indicator. *Scientometrics,* 29, 2, 271-297.

14 NEWMAN, M. E. J. (2001). The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci.* 98, 404-409

15 OKUBO, Y., MIQUEL, J. F., FRIGOLETTO, L. & DORÉ, J. C. (1992). Structure of International Collaboration in Science; Typology of Countries through Multi-variate Techniques using a Link Indicator. *Scientometrics,* 25, 321-351.

16 ROUSSEAU, R. (1992). Breakdown of the robustness property of Lotka's law: the case of adjusted counts for multi-authorship attribution. *Journal of the American Society for Information Science and Technology*, 43, 645-647.

17 PRICE, D. DE SOLLA (1963). *Little Science, Big Science*. New York: Columbia Univ. Press.

18 QIN, J. (1995). Collaboration and publication productivity: An experiment with a new variable in Lotka´s law. In *Proceedings of the 5th International Conference on Scientometrics and Informetrics, June 7-10, 1995, River Forest (USA).* M. E. D. Koenig & A. Bookstein (Ed.) (pp. 445-454). Medford: Learned Information.

19 TIJSSEN, R. J. W. & MOED, H. F. (1989). *Science and Technology Indicators.* Leiden: DSWO Press.

20 WINCH, R., KTSANES, T. & KTSANES, V. (1954). The Theory of Complementary Needs in Mate-Selection: An Analytical and Descriptive Study. *American Sociological Review,* 19, 241-249.

21 Wolf, Ch. (1996). *Gleich und gleich gesellt sich. Individuelle und strukturelle Einflüsse auf die Entstehung von Freundschaften*. Hamburg: Verlag Kovac.

22 Zitt, M., Bassecoulard, E. & Okubo Y. (2000). Shadows of the Past in International Cooperation: Collaboration profiles of the top five producers of science. *Scientometrics*, 47, 627-657.